CTF Challenge: Result Summary. Supplementary Material

Roberto Marabini[(a)], Bridget Carragher[(b)], Shaoxia Chen[(i)], James Chen[(m)], Anchi Cheng[(b)], Kenneth H. Downing[(d)], Joachim Frank[(e)], Robert A. Grassucci[(e)], J. Bernard Heymann[(l)], Wen Jiang[(f)], Slavica Jonic[(j)], Hstau Y. Liao[(e)], Steven J. Ludtke[(c)], Shail Patwari[(k)], Angela L. Piotrowski[(k)], Adrian Quintana[(g)], Carlos O.S. Sorzano[(g)], Henning Stahlberg[(h)], Javier Vargas[(g)], Neil R. Voss[(k)], Wah Chiu[(c)], Jose M. Carazo[(g)]

March 18, 2015

[(a)]Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain.

[(b)] The National Resource for Automated Molecular Microscopy, The Scripps Research Institute, La Jolla, CA 92037, USA

[(c)] Baylor College of Medicine, Houston, Texas 77030, USA

[(d)] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[(e)] Howard Hughes Medical Institute, Columbia University, NY 10032, USA

[(f)] Purdue University, Biological Sciences, IN 47907-2054, USA

[(g)] Biocomputing Unit, National Center for Biotechnology (CSIC), C/ Darwin, 3, Campus Universidad Autónoma, 28049 Cantoblanco, Madrid, Spain.

[(h)] Biozentrum, University of Basel, CH - 4058 Basel, Switzerland

[(i)] MRC-LMB, Cambridge CB2 0QH, UK. 01223 267000, United Kingdom

[j]IMPMC, Sorbonne Universités - CNRS UMR 7590, UPMC Univ Paris 6, MNHN, IRD UMR 206, 75005 Paris, France.

[k]Roosevelt University, Department of Biological, Chemical, and Physical Sciences, 1400 N. Roosevelt Blvd., Schaumburg, IL 60173, USA.

[l]Laboratory of Structural Biology Research, National Institute of Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892, USA

[m]Massachusetts Institute of Technology, USA

Corresponding Author: Roberto Marabini, Ph= 34 91 5854510, email: roberto@cnb.csic.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Cantoblanco
Madrid
Spain

Abstract

This work contains a detailed description of the the experimental data used in the "CTF Estimation Challenge" as well the analysis performed in order to establish the level of statistical significance at which the claim that "the results provided by two software packages are equal" can be rejected.

Description of Data Sets

Nine data sets were used in this Challenge, eight consisting of experimentally collected micrographs using a range of samples, microscopes and detectors, while the ninth data set was a collection of computer-simulated images.In order to provide a first estimate of how difficult the task was, we present:

- in Figure 1, a table containing: a representative image from each data set (column 3) and its corresponding power spectrum as a rotationally averaged plot (column 1) as well as a 2D representation (column 2).
- in Figure 2, the defocus estimation provided for all the Challenge participant for three representative micrographs. The first micrograph belongs to a challenging data set, for which the relative discrepancy among estimations was large (Figure 2a), the second to a data set for which the discrepancy was small (Figure 2b), and finally, the third micrograph belong to data set 9, which has been computer generated (Figure 2c).

Additionally, we also asked to the data providersthem for their own estimation of the CTF as well as the method they used to estimate it. This information is compiled in Appendix Supp-A.

[Figure 1 goes approximately here]

[Figure 2 goes approximately here]

Results: The Contributions to the CTF Challenge

An interesting, although complex question to be asked at this stage is whether the different software packages are equally good in estimating the CTF. When comparing results, we must bear in mind that not all participants have submitted estimations for all micrographs and that some contributions have not been provided by the package developers. In fact, we list in Table

1 the number of micrographs processed for each upload, noting that most of the uploads contain all 197 micrographs.

In order to determine if the means of two sets of data are significantly different from each other we used the Wilcoxon test.

Wilcoxon tests were computed for all pairs of uploads. Figures 3, 4, 5 and 6 show the result of performing this test when grouping the data in four different ways: (1) all experimental data sets (except for data set 8), (2) Pool 1, (3) Pool 2 and (4) the synthetic data set. In the following, and as it is the standard procedure in statistics, we will consider two uploads to be different if their corresponding p-value is smaller than 0.05.

All the figures present the results using a table in which the cell color is set to reflect the magnitude of the p-value obtained for that comparison. Values less than 0.05 are set to red. Values higher than 0.05 are set to blue. Additionally, cells that relate uploads that have not processed the whole data set under examination are set to black. In this way, in Figure 3, the row related with upload 287 is red for all cells, therefore, upload 287 is statistically different from any other upload (except, of course, when comparing with itself). On the other hand, the row related to upload 300 is blue for columns 337, 339 and 318, meaning that for these four uploads we cannot reject the hypothesis that the three are equivalent (Note: upload 340 p-value is smaller that 0.05, although it is difficult to establish it from the Figure).

A Protocols for Data Production and CTF estimation at the different data acquisition laboratories

Here we compile the different CTF protocols used at the different data acquisition laboratories.

Data Sets 1 & 2

Images of *GroEL* in 2μ hole C-flat grid were taken on an FEI F20 microscope operated at 200 kV. Leginon was used to reproduce the hysteresis of imaging sequence between calibration and data collection.

The images were taken in fixed sequence of pairs, or triplets (data set 1). The first of the image pair was taken at defocus values varying between 0.5, 1.0, 2.0, and 5.0μ and the second image (and third image in data set 1) was taken with the fixed nominal defocus of 1.5μ. After normal astigmatism adjustment on a carbon support area 1.5-2.0μ away from the target areas,

astigmatism was deliberately introduced by applying extra current to the X objective astigmator to induce 500-1000Å of difference in defocusU vs. defocusV.

Calibration

Identical imaging sequences were taken on carbon film at $62\frac{e}{Å^2}$. Parameters in $CTFFIND3$ were optimized for individual images to provide an estimate of the defocus difference in each pair (Δd), accounting for the hysteresis.

For the provided dataset, the fixed-defocus exposure CTF estimation was obtained using an over-exposed reference (defined in each dataset) using $CTFFIND3$. The harder-to-estimate mean defocus of the first of the pair of images was calculated from the calibrated defocus difference (Δd) and the offset provided by the difference between the fixed-defocus exposure and its nominal value, 1.5μ.

The amount of extra current applied in the data collection was pre-calibrated using a carbon film grid. However, the hysteresis of the astigmator and possible microscope alignment differences between calibration and data collection will affect the reproducibility of the astigmatism magnitude. Therefore, the data provider's value of defocusU and defocusV for the first exposure is only an average of the values obtained for the over-exposed reference.

Data Set 1

The images were taken on TIVPS F416 light-sensitive CMOS camera. Three exposures were taken at the same grid position. First exposure dose is $20\frac{e}{Å^2}$ taken at varying defocus. Second exposure dose is $20\frac{e}{Å^2}$ with fixed nominal defocus of 1.5μ. Third exposure image dose is $100\frac{e}{Å^2}$ taken at the same defocus as the second image. This third image serves as the reference image for the data provider's estimate of defocus and astigmatism and was not provided as part of the CTF Challenge data sets.

Data Set 2

The images were taken on Direct Electron DE-12 DDD camera. Two exposures were taken at the same grid position. First exposure dose is $20\frac{e}{Å^2}$. Second exposure dose is $100\frac{e}{Å^2}$. All 50 raw frames that

accompanied the second exposure were saved. The integrated exposure up to $20\frac{e}{\text{Å}^2}$ was used to create dataset 2. The integrated exposure up to $100\frac{e}{\text{Å}^2}$ serves as the reference for the data provider's estimate and was not provided as part of the CTF Challenge data sets.

Data Sets 3 & 4

Defocus pair images (each with a dose of $22\frac{e}{\text{Å}^2}$) were obtained on an FEI F30 Polara microscope operated at 300 kV acceleration voltage and 23,000x magnification, using the following protocol: Imaging was done using the Gatan K2 Summit DDD in dose fractionation (movie) mode. A series of 4 frames were collected at $5.5\frac{e}{\text{Å}^2}$ each. These frames were then aligned to correct for drift and the average of the aligned frames were used for further analysis.

Calibration

Defocus zero was approximated by focusing on a thick carbon area on a Quantifoil grid (Grid #1 with carbon). Minimal image contrast and appearance of a broad Gaussian-looking power spectrum were used as criterion to set zero defocus. Next, the microscope's defocus steps were calibrated by collecting a defocus series of images from 1 to 7μ in 1.0μ steps, still on the thick Quantifoil carbon. The defocus values of these images were calculated using the *defocus.spi* batch file written in SPIDER command language, along with *CTFMatch.py*, and these values were plotted on the Y axis versus the set defocus values on the X axis. This calibration ensured that we know exactly how to create a pair of micrographs with a given defocus difference. For the experimental datasets 3 and 4, defocus zero was approximated using low dose by focusing on adjacent thick carbon areas on a Quantifoil grid (GRID #2 and 3) which were approximately 2 microns away from the experimental area. Minimal image contrast and appearance of a broad Gaussian-looking power spectrum were used as criterion to set zero defocus.

Data Set 3

Grid #2 (with thin carbon, and with 60S particles over holes) was used to collect 12 defocus pairs in the experimental area of the Quantifoil grid, each with a defined nominal value and defocus increment

as determined in the calibration. For all 12 defocus pairs of this grid, defocus values were determined using *defocus.spi*, and defocus differences were verified in each case as matching the set increment in the experiment.

Data Set 4

Grid #3 (without carbon, and with 60S particles over holes) was used to collect 12 defocus pairs in the experimental area of the Quantifoil grid, each with a defined nominal value and defocus increment as determined in the calibration. For all 12 defocus pairs of this grid, defocus values were determined using *defocus.spi*, and defocus differences were verified in 9 cases as matching the set increment in the experiment. In the other 3 cases, the difference was off due to the fact that one or both micrographs was close to focus, making the estimation by *defocus.spi* fail. These three pairs were therefore left out from the data submitted to the challenge.

Data Set 5

The micrographs were recorded at 200keV with an electron dose of $16\frac{e}{\text{Å}^2}$, from apoferritin specimens on 1.2μ hole diameter Quantifoil grids. They were recorded on SO-163 film developed for 12 minutes in D19 developer. The nominal magnification was 59,000x. After calibration by cross-correlation against an atomic model of apoferritin from the PDB, the real magnification was found to be 61,400x. Each micrograph was recorded using nominal microscope defocus values of 1.0, 1.5, 2.0, 3.0, 5.0 μm. At each defocus, there are 3, 4, 3, 4, 3 micrographs respectively. The nominal defocus was adjusted by defocusing from zero before recording each image. The in-focus zero defocus level was calibrated by bisecting the mid-point of the minimal contrast position, using a live FFT on a TV-rate detector to observe the Thon rings from two diametrically opposed areas of the carbon film separated by 2.0 microns from the area shown in the images. After digitising the films on our in-house KZA scanner in 10 micron steps, the defocus and astigmatism of each image was determined using *CTFFIND3*, using a magnification of 61,400x and an amplitude contrast of 5%. The amount of possible astigmatism (100-200Å) and its direction showed no consistency from one image to the next, so we believe any astigmatism is minimal (less than 100Å).

Data Set 6

The specimen used was Equine Apo-Ferritin from Sigma-Aldrich cat# A3641, diluted with filtered 150μM NaCl. Quantifoil grid 1.3/1.2 copper 400 mesh was cleaned with ethyl acetate followed by MilliQ water rinse. Plasma clean shortly before applying the specimen aliquot. A Vitrobot Mark IV (FEI) was used for vitrification with the following parameters: holey carbon, 3μl specimen aliquot, 100% humidity, 22°C, 2 sec blot once. Images were recorded at 300 kV on a JEM 3200FSC electron microscope with the following characteristics: Cs = 4.1 mm, condenser aperture = 70 μm, objective aperture = 120 μm, specimen temperature = 87.2°K, in-Column Omega energy filter = 20 eV, microscope magnification = 40,000, detector magnification = 50,939, magnification calibration: Catalase crystal, $1.178\frac{\text{Å}}{pixel}$, Detector: DE-12 (Direct Electron), 6 μm/pixel, Specimen exposure: $20\frac{e}{\text{Å}^2}$.

Images were obtained as defocus pairs with various nominal defocus difference of 0.5, 1.0, 1.5, 2.0 μm etc. based on the pre-calibration of corresponding defocus between each focus knob (1,2,3 etc. clicks). The software used for assessment of defocus was $EMAN2$ *e2evalimage.py*.

Data Sets 7 & 8

Images were obtained focusing on a grid position about two micrometers adjacent to the exposure position. Focusing was done on the thicker carbon film of the Quantifoil grid. This was done by searching for the minimal contrast in the images, which are defined as "in focus" (defocus=0). Then, the objective lenses were defocused by a certain amount, and an image in the "exposure position" was recorded. Since the sample is on the thin carbon film that covered the entire grid, it is assumed that the physical height difference between focus position and exposure position should be small, no more than 100 nm. This offset will be the same for all images within one data set, since the images from each data set were recorded from the same grid. Calibration of the microscope nominal defocus was done by comparison with two programs, *2dx* and *CTFFIND*.

Data Set 9

Simulation follows the data model developed in Baxter et al. (2009) based on their experimental measurements of the SNRs for structural and post-CTF noise. The background structure (ice + thin carbon) was modeled by Gaussian noise, such that SNR = 1.4 at the object stage. A 4096 x

4096 object field was created by adding projection images of the 70S -EFG ribosome in random orientations and positions to the background structure. The resulting object field was Fourier-transformed, CTF-modified, and then inverse-Fourier transformed. The resulting simulated image was added to a Gaussian noise image of the same size, such that the resulting final simulated image has SNR = 0.06. The CTFs simulate the effect of the FEI Tecnai F30 Polara TEM operated in the bright-field mode at 300 kV, with Cs=2.26 mm, λ = 0.0196868Å, amplitude contrast ratio = 0.1, pixel size 1.525 Å, and Gaussian envelope 10,000.

Copyright Notice

Permission to use the data sets described in this section has been granted by their authors subject to the terms of the copyright available at `http://i2pc.cnb.csic.es/3dembenchmark/LoadCtfInformation.htm`. In particular, this paragraph should be quoted in any work that uses the data provided in this Challenge.

> "This work has made use of electron micrographs provided by the following researchers: Richard Henderson (Medical Research Council, UK), Henning Stahlberg (University of Basel), Joachim Frank (Columbia University), Wah Chiu (Baylor College of Medicine), Anchi Cheng (Scripps Research Institute), as well as resources from the Biocomputing Unit of the Spanish National Center for Biotechnology (CNB-CSIC) as part of an ESFRI Instruct support project from the Ministry of Economy and Competitiveness (AIC-A-2011-0638)".

Notice: For any reuse or distribution, you must include this copyright notice.

B Participants' Comments

This appendix collects the comments from the challenge participants on the performance of their particular contributions. The opinions in this section express the participant's personal view and have not been agreed on among the rest of the article authors.

The detailed individual results are made available at URL `http://i2pc.cnb.csic.es/3dembenchmark/LoadAnalyze.htm?subtaskId=2`

B.1 Xmipp: uploads 282, 291 and 298

Xmipp was used in three uploads made by different people. After automatic CTF estimation, Xmipp allows a manual correction. The three participants were asked to be more (upload 291) or less (upload 298) strict with the criteria for CTF manual adjustment.

- The first participant (upload 282) was told to behave as a typical expert user, that is, to examine the CTF estimation and recalculate it in those cases in which visual inspection shown differences between the estimation and the experimental PSD.
- The second participant (upload 291) followed a more strict policy regarding the estimation of the CTF.
- Finally, the third participant (upload 298) followed a less strict policy regarding the estimation of the CTF.

Our first comment is that the uploads are all very similar, indicating the robustness of the algorithm. In terms of performance, Xmipp (upload 282) is among the more accurate methods. In particular, is the third one for the more difficult data sets (Pool 2) and for the group containing data sets 1, 2, 3, 4, 5, 6 and 7.

B.2 Particle: upload 299

The results in this submission were processed by the CTF-module of the PARTICLE package (`www.image-analysis.net/EM`) for single-particle EM data analysis and 3D reconstruction. The benchmarks of the Challenge, particularly the synthetic data (Dataset 9) that comes with the ground-truth, provide an objective test of the functionality of this CTF module. The detail of the data analysis and results from the PARTICLE package is online at URL `http://www.image-analysis.net/Challenge/CTF2013`.

On the synthetic data set, PARTICLE is able to resolve the defocus parameters within 10nm in the amplitudes and much less than 1-degree (see Supplementary Material) in astigmatic angles. On the experimental data sets, the PARTICLE results closely track the "consensus value". It is important to note that 1) a good performance on the synthetic benchmark is necessary yet insufficient for the method validation; and 2) the consensus estimates on experimental data do not represent the truth and the error is unknown. The ultimate evaluation of any CTF determination method will

be the resolution of 3D reconstructions. Hopefully that will be in the future plan of the Challenge

B.3 Appion: upload 310

This is an upload made by a novice user of CTF methods. This submission represents the best available result of running several methods within the Appion pipeline (Lander et al., 2009). Note that several other submissions including #300, #301, and #304, originating from the group headed by Neil Voss, reflect the use of individual methods within Appion. For submission #310, Appion 3.0 was used to automatically choose the optimal CTF estimation after comparing all CTF estimation trials available within the pipeline. Appion was used to launch CTF estimates using Ace1, Ace2 (Mallick et al., 2005), and CTFFind (Mindell and Grigorieff, 2003) on each data set using default values provided by the Appion interface and using various image binning values as advised by the user guide. Thus for each data set, 8 estimation runs were launched: one CTFFind, three Ace2 with binning at 1, 2, and 4, and four Ace1 with binning at 0.5, 1.0, 1.5, and 2.0. The optimal results were then automatically selected by Appion based on a quality assessment method developed by Neil Voss (publication under review). Ace 1 was optimal for 24% of the images, Ace 2 for 41% of the images, and CTFFind for 35% of the images. Overall, for the group of more difficult data sets, the Appion #310 novice submission provided results that were similar to the CTFFind expert submission (#287), at least as based on the RES-90 as defined by this paper. Note that submission #287 for CTFFind used a newer and improved version of the CTFFind software than was available in the Appion pipeline.

Based on the Appion submission #310 we conclude that:

- Currently, no single method provides optimal results for all images; multiple methods should be used to achieve the best results.
- Considering that Appion #310 results are generally close to "Consensus", the method Appion uses for CTF estimation quality assessment appears to be effective.
- Appion #310 as a novice level submission provided better overall RES-90 values than most of the Developer and Expert level submissions. This is likely because it supports running multiple methods, including CTFFind, Ace1, and Ace2, and automatically chooses the optimal results using an unbiased assessment method.

B.4 Bsoft: upload 312

Bsoft version 1.8.8 was used by the developer. All data sets were analyzed automatically with no user intervention, to be able to test the automatic fitting functions in Bsoft. I used the following command line:

```
bctf -v 1 -act prepfit -resol 50,5 -out set_003_fit.emx set_003.emx
```

After seeing the comparative results, I redid the CTF fits manually to identify problem areas. These fall into the following categories:

1. Poor quality micrographs: These have power spectra where the CTF cannot be interpreted by any means.

2. Difficult micrographs: The oscillatory nature of the CTF curve means that an algorithm may find multiple potential solutions and pick the wrong one. With low SNR (such as those close to focus), the radial power spectrum is confusing even to a human being.

3. The astigmatism detection algorithm in Bsoft is based on an assumption that the variance within a range of frequencies will be a maximum at the correct parameters. This may also suffer from a multiple solution problem and needs to be re-examined.

B.5 ACE2, Appion Interactive and Phasor: uploads 300, 301 and 304

For the CTF challenge, Dr Voss' lab submitted three estimation sets using three different programs. The first set, "#300 ACE2 Expert final submission," used some of the more advanced features of the ACE2 program. ACE2 is based on the ACE1 program (Mallick et al., 2005), but was written to be more portable, faster, and provide reliable astigmatism estimation. The second set, "#301 Appion Interactive CTF Submission" was using an interactive display to determine the CTF. *Interactive CTF* is a manual CTF estimation program with buttons to streamline the CTF fit process. The third and final set, "#304 Phasor CTF dataset" was created using a new program called Phasor CTF that uses an experimental least squares refinement method. For all three of our datasets, the programs were run multiple times and the optimal estimation parameters were chosen using a novel CTF resolution method (Sheth et al. submitted).

One caveat of this challenge is that the amplitude contrast was not included in the submissions. The amplitude contrast can affect the defocus estimate especially for images of poor CTF quality. Some CTF programs use a fixed amplitude contrast, whereas all of our methods aggressively attempt

to optimize the amplitude contrast. Programs that allow the amplitude contrast to vary will be at a disadvantage when the best defocus estimate is taken to be the consensus from many fixed amplitude contrast submissions.

C Authors' Contribution

This work has made use of electron micrographs provided by the following researchers: A. Cheng (Scripps Research Institute); W. Chiu & J. Jakana (Baylor College of Medicine); J. Frank & R. A. Grassucci & H. Y. Liao (Columbia University); R. Henderson & S. Chen (Medical Research Council); H. Stahlberg & M. Chami & K. Goldie (University of Basel); as well as resources from the Biocomputing Unit of the Spanish National Center for Biotechnology (CNB-CSIC) as part of an ESFRI Instruct support project from the Ministry of Economy and Competitiveness (AIC-A-2011-0638)

The individuals and institutions that have participated in the Challenge are summarized in Table 2:

[Table 2 goes here]

Data Analysis was initially performed mainly by: J.M. Carazo (CNB-CSIC)), W. Chiu (Baylor College of Medicine), K. Downing (Lawrence Berkeley National Laboratory), W. Jiang (Purdue University), S. Ludke (Baylor College of Medicine), R. Marabini (UAM) and C.O.S. Sorzano (CNB-CSIC). The initial analysis was further refined though iteration with all data providers.

References

Baxter, W. T., Grassucci, R. A., Gao, H., Frank, J., May 2009. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. J. Struct. Biol. 166 (2), 126–132.

Lander, G. C., Stagg, S. M., Voss, N. R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P. W., Lyumkis, D., Potter, C. S., Carragher, B., Apr 2009. Appion: an integrated, database-driven pipeline to facilitate EM image processing. J. Struct. Biol. 166 (1), 95–102.

Mallick, S. P., Carragher, B., Potter, C. S., Kriegman, D. J., Aug 2005. ACE: automated CTF estimation. Ultramicroscopy 104 (1), 8–29.

Mindell, J. A., Grigorieff, N., 2003. Accurate determination of local defocus and specimen tilt in electron microscopy. J. Struct. Biol. 142 (3), 334–347.

Table Caption List

Table 1: Uploads description in terms of processed micrographs per upload. The total number of micrographs is 197. For those cases in which a dataset has been partially processed, we have followed the notation Y(X) + comment, where Y refers to the dataset, X refers to the number of processed micrographs and comment describes which datasets were either processed or not processed.

Table 2: Summary of the participant contribution to the CTF Challenge. Note: Upload 338 corresponds to the consensus value

Figure Caption List

Fig. 1: Examples of representative Power Spectral Densities and micrographs for the different data sets. A radial profile is presented on the left hand column and a 2D image on the center, in logarithmic scale in both cases. In order to increase contrast, all frequencies smaller than 0.8 (that is, 10 pixels) have been masked out. Right column shows the original micrograph. Note that a downsampling factor of two has been applied to all micrographs before processing, so as to obtain a zoom into the central part of the spectrum. Micrographs have been selected so that they have an average defocus as close as possible to 1.8 μ m.

Fig. 2: Scatter plots showing the CTF parameter estimations for three representative micrographs belonging to datasets 1, 3 and 9, respectively. The first micrograph comes from a challenging data set, for which the relative discrepancy among estimations was large (a). In turn, the second micrograph belongs to a data set for which the discrepancy was smaller (b). Finally, dataset 9 is formed by the computer generated images (c). Color bar shows astigmatism angle. Note that x and y-axis ranges are different in the different plots

Fig. 3: Wilcoxon test computed for data sets 1, 2, 3, 4, 5, 6, 7. Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have

not processed the whole data set under examination are set to black.

Fig. 4: Wilcoxon test computed for data sets 3, 4, 5 and 7 considered as the ones with the smaller discrepancies. Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have not processed the whole data set under examination are set to black.

Fig. 5: Wilcoxon test computed for data sets 1, 2 and 6 considered the ones with larger discrepancies. Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have not processed the whole data set under examination are set to black.

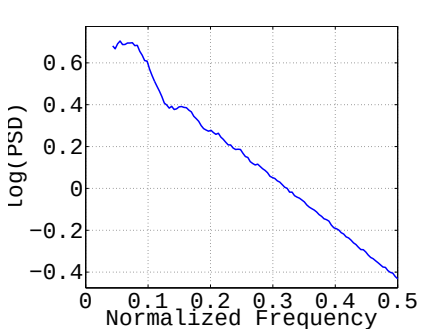
Fig. 6: Wilcoxon test executed for data set 9 (synthetic one). Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have not processed the whole data set under examination are set to black.

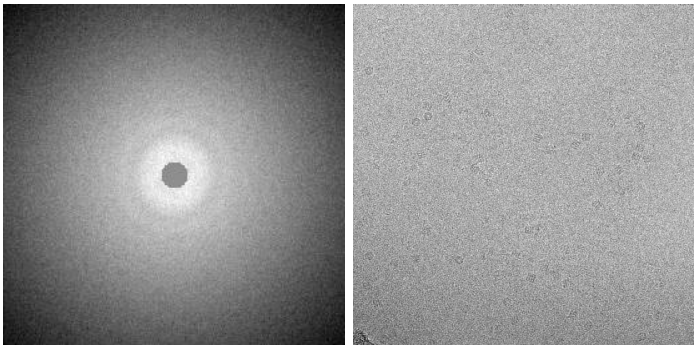
Upload Number	# Micrographs Downloaded	Comments	package
282	197		xmipp
287	197		ctffind
291	197		xmipp
292	197		sparkx
296	197		fitctf2
298	163	Results for dataset 6 are not included	xmipp
299	197		particle
300	197		ace
301	197		appion
303	16	Only dataset 1 was processed	eman
304	197		ace-appion
310	197		appion
312	197		bsoft
314	197		fei
318	197		spider
336	59	Only data sets 5, 6, 9 were processed	eman
337	197		dudelft
338	197		consensus
339	188	Astigmatism angle was disabled. Datasets missing: 8(1) and 9	e2rawdata (eman.2.1)
340	187	Astigmatism angle was disabled. Datasets missing: 8(2) and 9	e2ctf (eman.2.1)
341	107	Applied to data sets with astigmatic information. Datasets missing: 1, 2, 7, 8	e2ctf (eman.2.1)
344	172	Some images missing from dataset: 1 (3), 2 (6), 8 (16).	imagic

Table 1: Uploads description in terms of processed micrographs per upload. The total number of micrographs is 197. For those cases in which a dataset has been partially processed, we have followed the notation Y(X) + comment, where Y refers to the dataset, X refers to the number of processed micrographs and comment describes which datasets were either processed or not processed.

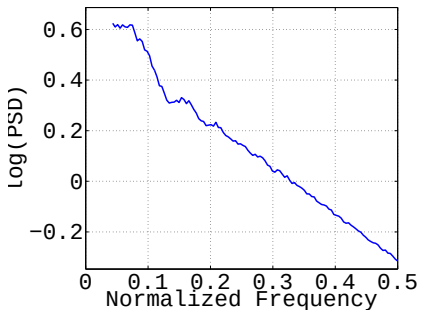
Upload	Participant	Affiliation
282	J. Vargas	CNB-CSIC
287	N. Grigorieff	Brandeis University
291	C.O.S. Sorzano	CNB-CSIC
292	R. Efremov	Max Planck Institute for Molecular Physiology
296	R. Yan	Purdue University
298	S. Jonic	CNRS
299	J. Chen	Massachusetts Institute of Technology
300	N. Voss	Roosevelt University
301	N. Voss	Roosevelt University
303	X. Huang	Institute of Biophysics Chinese Academy of Sciences
304	N. Voss	Roosevelt University
310	A. Herold	The Scripps Research Institute
312	B. Heymann	NIH
314	E. Franken	FEI Company
318	R. Langlois	Columbia University
336	L. Kong	Institute of Biochemistry and Cell Biology Chinese Academy of Sciences
337	M. Vulovic	TU Delft/ LUMC
338	R. Marabini	UAM-Spain
339	S. Ludtke	Baylor College of Medicine
340	S. Ludtke	Baylor College of Medicine
341	S. Ludtke	Baylor College of Medicine
344	R. Righetto	LNNano/Unicamp

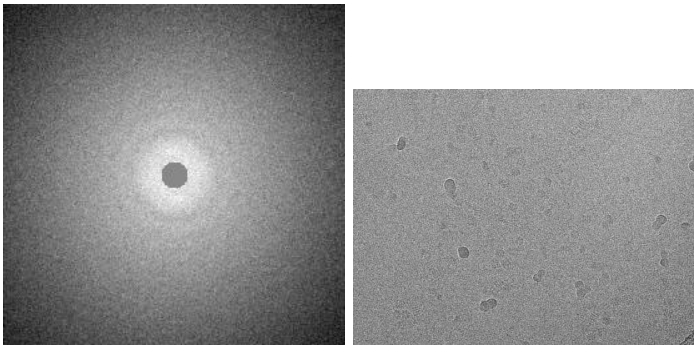
Table 2: Summary of the participant contribution to the CTF Challenge. Note: Upload 338 corresponds to the consensus value



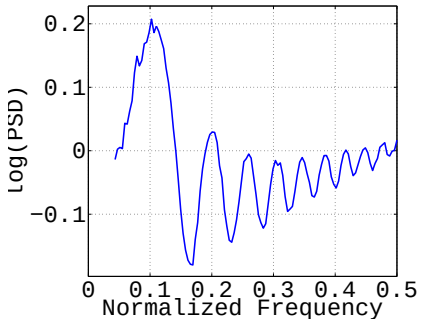


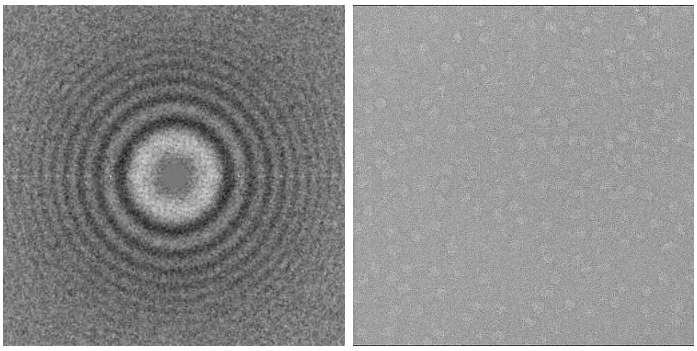
(a) Data set 1



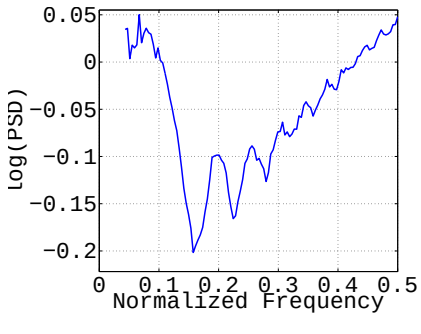


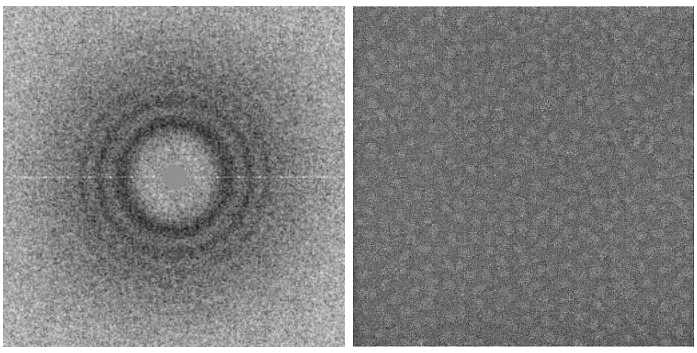
(b) Data set 2



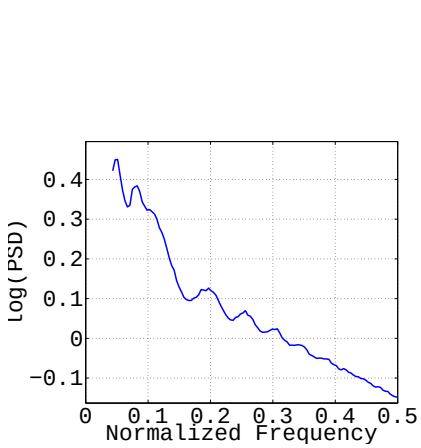


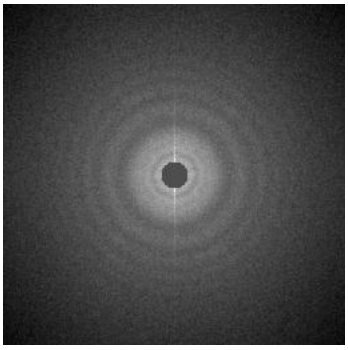
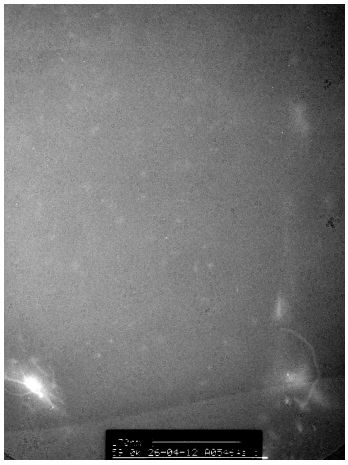
(c) Data set 3



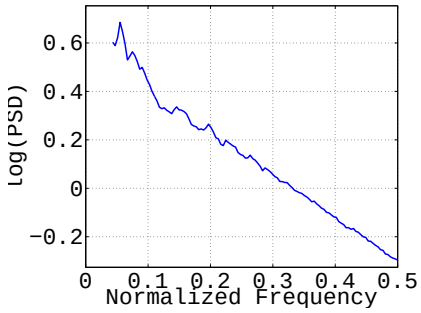


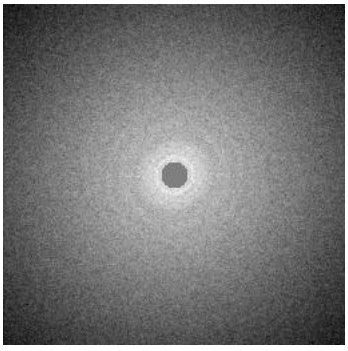
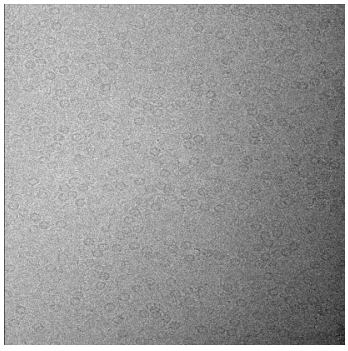
(d) Data set 4



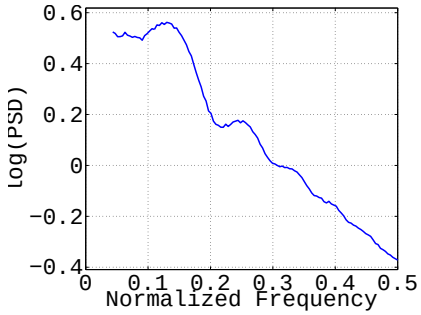



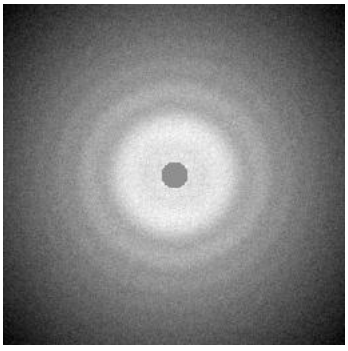
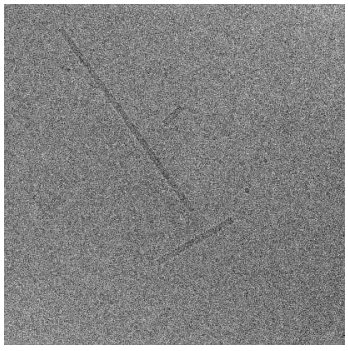
(e) Data set 5



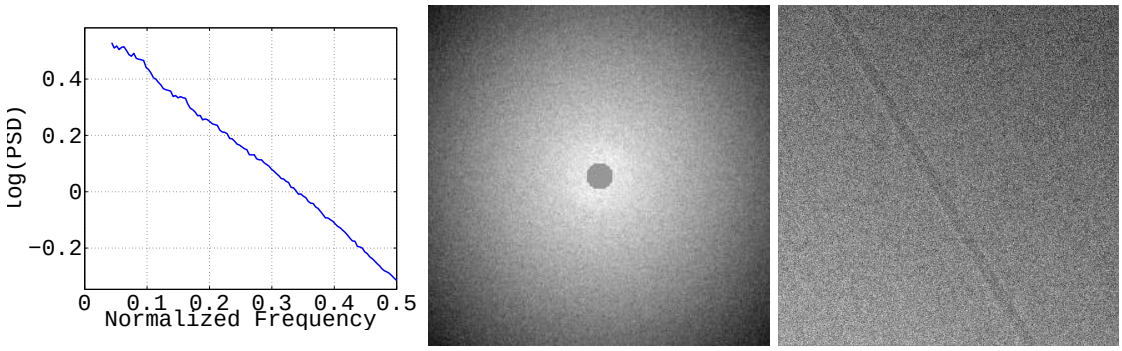



(f) Data set 6

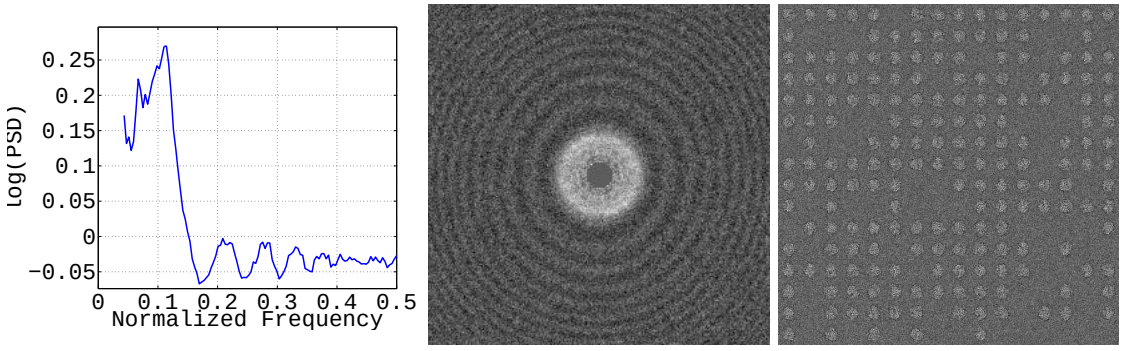


(g) Data set 7

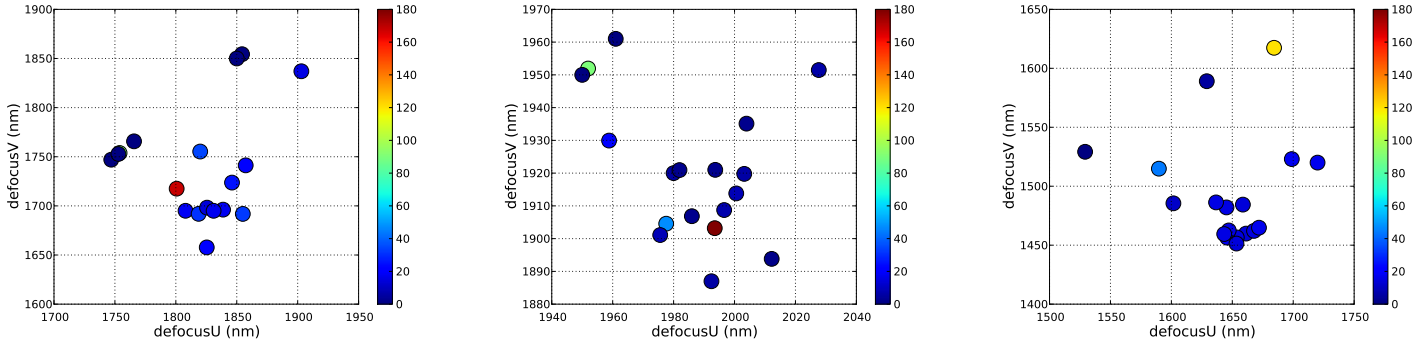


(h) Data set 8



(i) Data set 9

Figure 1: Examples of representative Power Spectral Densities and micrographs for the different data sets. A radial profile is presented on the left hand column and a 2D image on the center, in logarithmic scale in both cases. In order to increase contrast, all frequencies smaller than 0.8 (that is, 10 pixels) have been masked out. Right column shows the original micrograph. Note that a downsampling factor of two has been applied to all micrographs before processing, so as to obtain a zoom into the central part of the spectrum. Micrographs have been selected so that they have an average defocus as close as possible to 1.8 μ m.



(a) Dataset 1 Image 4 (b) Dataset 3 Image 6 (c) Dataset 9 Image 2

Figure 2: Scatter plots showing the CTF parameter estimations for three representative micrographs belonging to datasets 1, 3 and 9, respectively. The first micrograph comes from a challenging data set, for which the relative discrepancy among estimations was large (a). In turn, the second micrograph belongs to a data set for which the discrepancy was smaller (b). Finally, dataset 9 is formed by the computer generated images (c). Color bar shows astigmatism angle. Note that x and y-axis ranges are different in the different plots

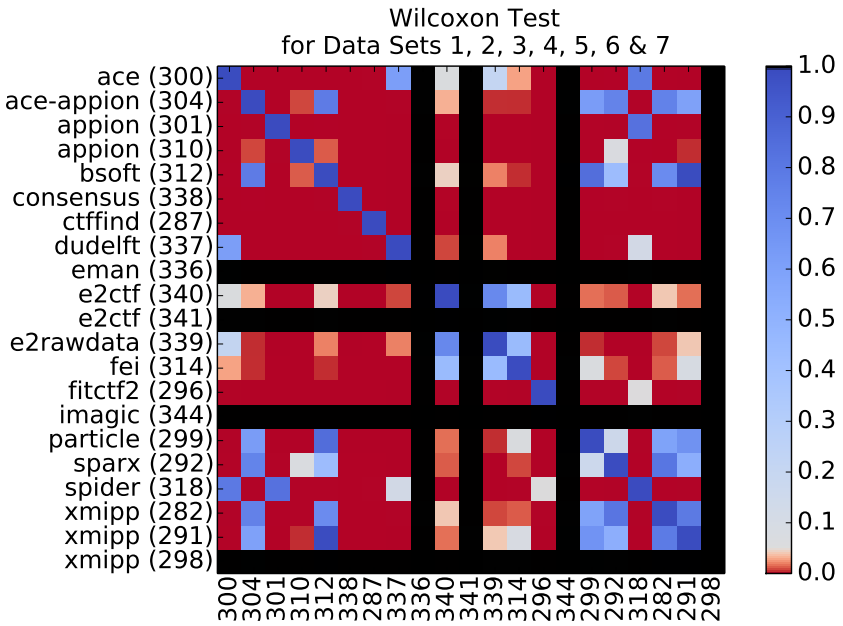


Figure 3: Wilcoxon test computed for data sets 1, 2, 3, 4, 5, 6, 7. Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have not processed the whole data set under examination are set to black.

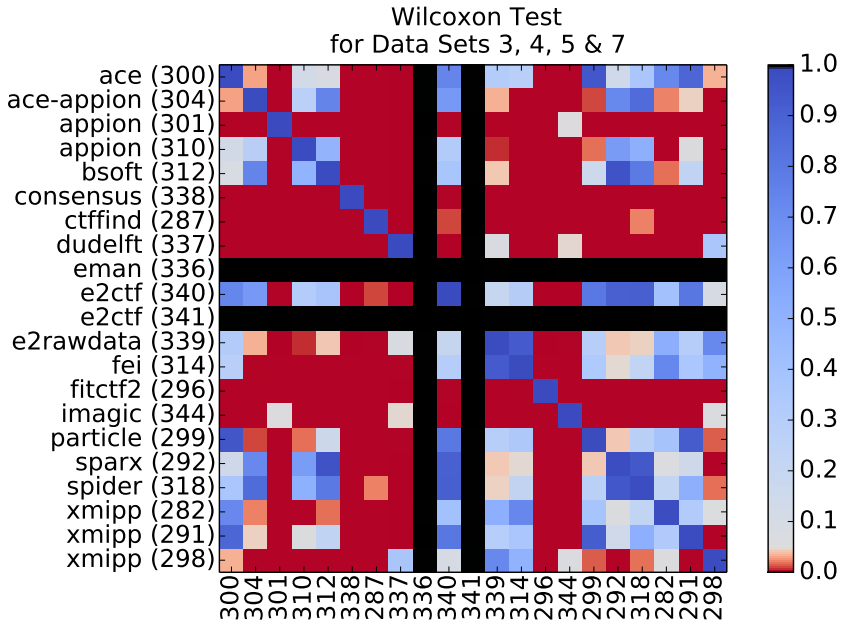


Figure 4: Wilcoxon test computed for data sets 3, 4, 5 and 7 considered as the ones with the smaller discrepancies. Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have not processed the whole data set under examination are set to black.

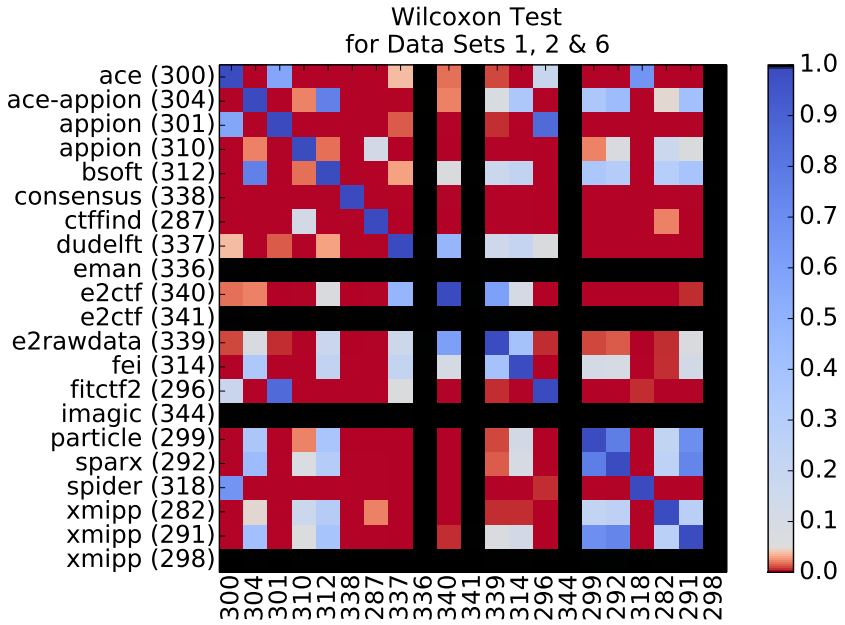


Figure 5: Wilcoxon test computed for data sets 1, 2 and 6 considered the ones with larger discrepancies. Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have not processed the whole data set under examination are set to black.

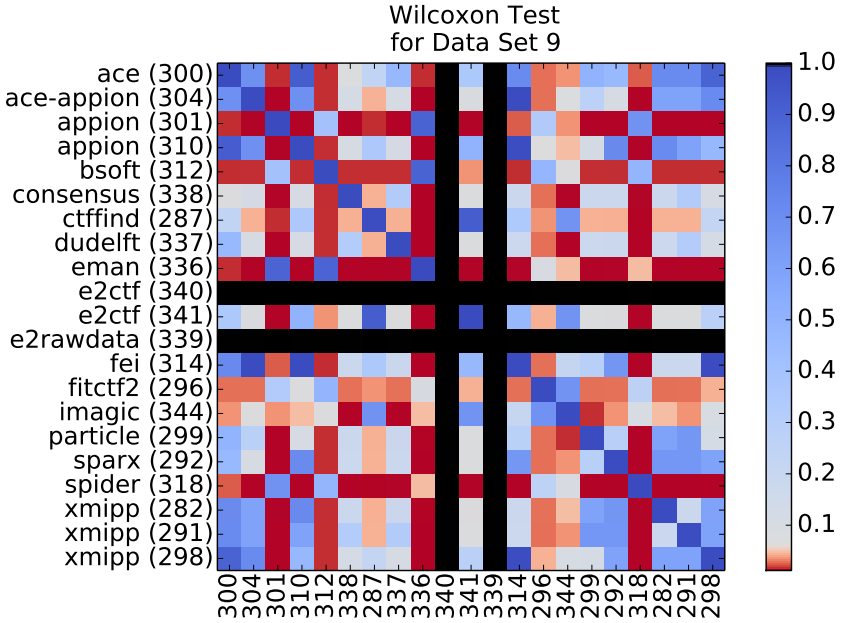


Figure 6: Wilcoxon test executed for data set 9 (synthetic one). Values less than 0.05 (in red) reject the null hypothesis that both population are indistinguishable (or, to be precise, that the difference population is symmetric and the median is zero valued). Cells that relate uploads that have not processed the whole data set under examination are set to black.